Large-scale image classification using ensembles of nested dichotomies

Arnau RAMISA [a,1] and Carme TORRAS [a]
[a] *Institut de Robòtica i Informàtica Industrial, CSIC-UPC, Spain*

Abstract. Many techniques to reduce the cost at test time in large-scale problems involve a hierarchical organization of classifiers, but are either too expensive to learn or degrade the classification performance. Conversely, in this work we show that using ensembles of randomized hierarchical decompositions of the original problem can both improve the accuracy and reduce the computational complexity at test time. The proposed method is evaluated in the ImageNet Large Scale Visual Recognition Challenge'10, with promising results.

Keywords. large-scale image classification, classifier ensembles, ensembles of nested dichotomies

Introduction

Most state-of-the-art methods for large-scale image classification use (compressed) high dimensional representations together with linear binary classifiers arranged under a One-versus-Rest (OvR) strategy [1]. With this approach, computational complexity at test time is linear in the number of classes, which may be a bottleneck with very large numbers. In order to attain sub-linear complexity in the number of classes at test time, it is possible to organize classifiers in a tree or DAG that allows using a branch and bound strategy, so that only a relevant subset of the classifiers are evaluated. However, these methods usually come at the cost of a more complex training procedure or a loss in accuracy. Another possibility is to generate tree structures with an inexpensive method (e.g. randomly) to partition the classes, and use ensembles or forests of these trees to maintain, or improve, the overall classification accuracy. This type of approaches are generally considered to be more expensive than single trees at test time, but they are very simple to implement and train, and can lead to improvements in accuracy with respect to OvR if a forest of sufficient size can be afforded.

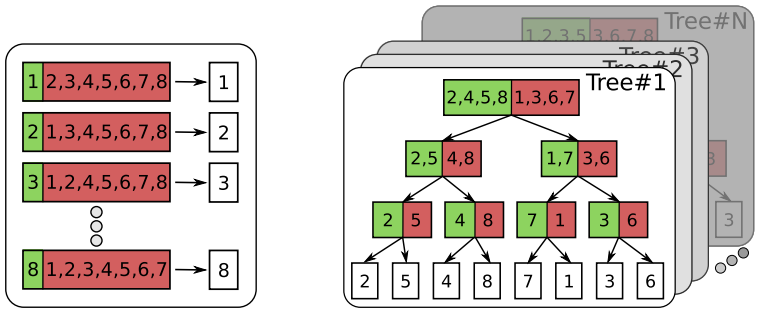
Moreover, in this work we will show that, when using a simple top-down traversal of the trees at test time, the Ensembles of Class-Balanced Nested Dichotomies (ECBND) approach proposed by Dong et al. [2] can lead to lower computational complexity at test time *and* improved results in the ImageNet Large Scale Visual Recognition Challenge'10 (LSVRC'10) dataset[2] when compared to the commonly used OvR strategy.

1. Ensembles of nested dichotomies

In this section, we briefly review the method proposed by Dong et al. [2], that we have selected for our large-scale image classification task. Let $\mathcal{C} = \{C_1, C_2, \cdots, C_N\}$ be the set

[1] Corresponding Author: Arnau Ramisa, Institut de Robòtica i Informàtica Industrial, CSIC-UPC. Parc Tecnològic de Barcelona. C/ Llorens i Artigas 4-6. 08028 Barcelona. E-mail: aramisa@iri.upc.edu.

[2] `http://www.image-net.org/challenges/LSVRC/2010/`



(a) One-versus-Rest (b) Class-Balanced Nested Dichotomies

Figure 1. Schema of the OvR and the ECBND approaches. Red and green boxes represent classifiers (the latter for the positive classes and the former for the negative ones), and the white boxes represent the probability of a class. At test time, OvR predicts the class with the maximum probability, and ECBND the class with the highest average probability in the forest.

of classes of our classification problem. Nested dichotomies are binary trees constructed as follows: starting with the set of all classes in the root node, recursively (randomly) splitting the set of classes that reach a node in two mutually exclusive groups (V_R and V_L), until leaves with a single class are attained. Then, at each node, a classifier is trained using the samples from classes in V_R as positives and samples from classes in V_L as negatives. As in [3], we use logistic regression models as node classifiers.

To compute the predicted class for a new test example x, we consider two approaches: The first approach, named *Full Branch Aggregation* along this work, is the one used in Dong et al. [2], and uses the fact that the sub-problems are statistically independent. Conditional probability estimates $p(c \in V_{i,R}|x, c \in V_i)$ and $p(c \in V_{i,L}|x, c \in V_i)$ are obtained at each node i of the tree, and the class probability estimates for the complete problem are computed as the product of the conditional probabilities over all the internal nodes of the tree:

$$p(c = C_j|x) = \prod_{i=1}^{N-1} I(c \in V_{i,R})p(c \in V_{i,R}|x, c \in V_i) + I(c \in V_{i,L})p(c \in V_{i,L}|x, c \in V_i), \quad (1)$$

where $C_j \in \mathcal{C}$ is the class we want to compute the probability for, and $I(b)$ is the indicator function, that outputs one in case the argument b is true, and zero if it is false.

The second approach, named *Single Branching* for future reference, consists of a simple top-down traversal of the tree that starts from the root node, recursively computes the prediction for a test example using the classifier at the current node, and descends via the right child if the test example is classified as positive and via the left child otherwise; thus, contrarily to *Full Branch Aggregation*, only a single branch has to be evaluated. Finally, when a leaf is reached, the associated class is predicted with probability one.

Since, a priori, all possible nested dichotomies that we can generate are equally likely and independent, it is possible to combine the output of many such nested dichotomies to obtain more robust class probability estimates. Collections of trees, constructed by sampling randomly and with replacement from the space of possible dichotomies, are called Ensembles of Nested Dichotomies (END). Given an ensemble of trees, we can compute the probability of each class in the original set as the average prob-

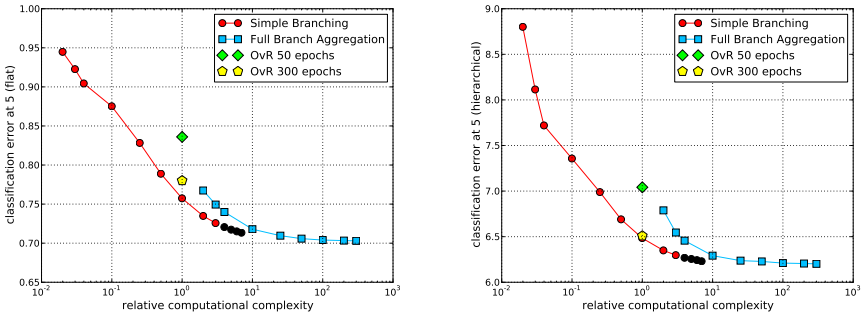


Figure 2. Results of the ECBND and OvR approaches on the LSVRC'10 dataset. The graph in the left corresponds to the *flat* error measure, and the one on the right corresponds to the *hierarchical* error measure proposed for this dataset. The markers in the ECBND curves correspond to ensembles of sizes 2, 3, 4, 10, 25, 50, 100, 200 and 300, respectively. We extended the ensemble for the *Single Branching* experiment to 400, 500, 600 and 700 trees to better determine where it saturates.

ability for the class in all trees. Then, we can define the final prediction of the ensemble as the class that maximizes the computed probabilities.

At this point we can focus on ensembles restricted to a particular subset of trees where, at each node, classes are split into two equal-sized subsets, since these are guaranteed to be the most efficient; i.e. they will only have depth $log_2(N)$. Dong et al. [2] called them Ensembles of Class-Balanced Nested Dichotomies (ECBND). Figure 1 shows a graphical representation of OvR and ECBND approaches.

2. Experimental results

We have evaluated the Ensembles of Class-Balanced Nested Dichotomies (ECBND) approach on the ImageNet Large Scale Visual Recognition Challenge'10 (LSVRC'10) dataset. This dataset comprises approximately 1 million training and 150 thousand testing images, assigned to 1000 different classes that span a significant portion of the ImageNet hierarchy, with categories as diverse as "in-line skate", "tiger", "geyser", "oak tree" or "restaurant". In order to account for images containing more than one of the objects in the dataset, the evaluation criterion recommended for this dataset is the error at five, i.e. five predictions are allowed for each test image. Note that, although we are not using it here, this robust evaluation measure could be directly optimized with a structured loss, as in McAuley et al. [4].

In order to quickly assess the potential of the proposed approach and to facilitate the reproducibility of our results, we used a Bag of Features (BoF) representation of size 1000 built on the demonstration features that come with the LSVRC'10 dataset. Finally, we applied the Hellinger's kernel explicit embedding [5] to the BoF vectors.

To deal with the computation and memory requirements of this dataset, we have used logistic regression models trained with Stochastic Gradient Descent (SGD) [6], as it is common in state-of-the-art large-scale image classification methods [7]. The parameters of the training algorithm were cross-validated on a small subset of the data, and then used in the rest of the experiments. In particular, we selected 50 training epochs as a good compromise between training time and accuracy.

In Figure 2 we can see the accuracy obtained with the Dong et al. [2] method variants and with the traditionally used OvR classifier as a function of the computational complexity at test time, taking that of the one-versus-rest classifier as unity. As can be observed, the *Single Branching* method is more accurate at the same complexity, and faster at the same accuracy than equivalent OvR classifiers, and as accurate as OvR classifiers that went through six times more training epochs and are close to asymptotic behavior. When *Full Branch Aggregation* is used, results are overall more accurate, but at the expense of more testing time. Interestingly, *Single Branching* is also more accurate than *Full Branch Aggregation* at the same computational complexity. This result suggests that increasing the size of the forest is more beneficial than exploiting the trees better.

3. Conclusions

In this work we propose to use an ensemble of nested dichotomies to both improve the accuracy and reduce the computational cost, at test time, in large-scale multi-class image classification problems. In particular, we adopted the Ensembles of Class-Balanced Nested Dichotomies (ECBND) proposed by Dong et al. [2], and evaluated direct top-down traversal of the tree to compute the prediction for every tree that scales well to large numbers of classes. We have found that this simpler method is able to obtain better accuracy than OvR classifiers, and attain the same accuracy at lower computational cost than both the OvR classifiers and the original ECBND method, which obtains the overall lowest error at a higher cost.

Typically, hierarchical decompositions of classification problems are either very computationally expensive to train, or degrade performance with respect to one-versus-rest. We believe that the proposed approach opens a way for applying such methods in large-scale image classification.

Acknowledgements

This work was supported by the EU project IntellAct FP7-269959, and by the Catalan Research Commission through SGR-00155. A. Ramisa worked under a JAE-Doc grant from CSIC and FSE.

References

[1] F. Perronnin, Z. Akata, Z. Harchaoui, and C. Schmid, "Towards good practice in large-scale learning for image classification," *2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3482–3489, June 2012.

[2] L. Dong, E. Frank, and S. Kramer, "Ensembles of balanced nested dichotomies for multi-class problems," in *Knowledge Discovery in Databases: PKDD*, pp. 84–95, 2005.

[3] E. Frank and S. Kramer, "Ensembles of nested dichotomies for multi-class problems," *Twenty-first International Conference on Machine Learning (ICML)*, p. 39, 2004.

[4] J. J. McAuley, A. Ramisa, and T. S. Caetano, "Optimization of Robust Loss Functions for Weakly-Labeled Image Taxonomies," *International Journal of Computer Vision*, pp. 1–19, Sept. 2012.

[5] F. Perronnin, J. Sánchez, and Y. Liu, "Large-scale image categorization with explicit data embedding," in *2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2297–2304, 2010.

[6] L. Bottou and O. Bousquet, "The Tradeoffs of Large Scale Learning," in *Advances in Neural Information Processing Systems*, 2007.

[7] F. Perronnin, S. Jorge, and T. Mensink, "Improving the Fisher Kernel for Large-Scale Image Classification," in *European Conference on Computer Vision*, vol. 6314, pp. 143–156, 2010.